

SMS text similarity calculation based on topic model

Chengfang Tan^{1, 2*}, Caiyin Wang², Lin Cui^{1, 2}

¹School of Information Engineering, Suzhou University, Suzhou 234000, Anhui, China

²Intelligent Information Processing Lab, Suzhou University, Suzhou 234000, Anhui, China

Received 1 August 2014, www.cmnt.lv

Abstract

The traditional text similarity calculation is mainly based on the statistical method and the semantic method, it exists data sparse and high-dimensional problems and so on. In order to improve the ability of SMS text similarity calculation, this paper puts forward a kind of similarity calculation method based on topic model. By using LDA (Latent Dirichlet Allocation) to model SMS document set and inference parameter via Gibbs sampling algorithm. The topic-word probability distribution and document - topic probability distribution of the SMS document set are generated. Then use JS (Jensen-Shannon) distance formula to calculate SMS text similarity, finally perform the text clustering experiments on the similarity matrix by single-pass incremental clustering algorithms. Compared with traditional text similarity calculation method, experimental results show that this proposed method can obtain better F-measure, which proves the effectiveness and superiority of the proposed text similarity calculation method.

Keywords: SMS text, similarity calculation, topic model, text clustering, latent Dirichlet allocation

1 Introduction

Text similarity calculation plays an important role in the field of information processing, it is one of the important factors to achieve effective text clustering, and it has been widely used in the field of information retrieval, text copy detection, Q & A system and so on.

The existed text similarity calculation method can be divided into two categories: the statistical method based on TF-IDF and the semantic method based on dictionary [1]. The statistical method based on TF-IDF (Term Frequency - Inverse Document Frequency) represents the text as vector space model and calculates the cosine between the vectors to get the size of text similarity. Its disadvantage is that this method needs large-scale corpus support, ignores the existence of semantic relationships between words, and text representation model has high dimensionality and sparse. The semantic method based on dictionary often represents text as word frequency vector model, constructs semantic relations between words by using domain-specific knowledge base, such as WordNet, HowNet and other external dictionaries. Compared with the statistical method based on TF-IDF, this method does not require the support of large-scale corpus, and has higher accuracy, but the establishment of knowledge base is a complicated project, furthermore, it is difficult to solve the semantic problem of words which are not logged in the dictionary, and it does not take into account the dimension reduction, lacks the definition of text similarity measure.

In this paper, we propose the SMS text similarity calculation method based on topic model, which makes full use of modelling advantages of topic model, maps SMS text corpus to each topic space and mines

relationships hidden in the text between different topics and words. We use LDA to model SMS text, and then obtain the topic-word probability distribution and document- topic probability distribution from the model, based on this calculate the SMS text similarity by using famous JS distance formula. Experimental results show that this method has better F-measure.

2 Theory of topic model

LDA model is the typical representative of statistical topic model. Because of its superiority in the text modelling, it becomes a research hotspot in recent years. LDA was put forward by Blei et al [2]. It is a three-tier structure that contains words, topics and documents. The structure of LDA is shown in Figure 1.

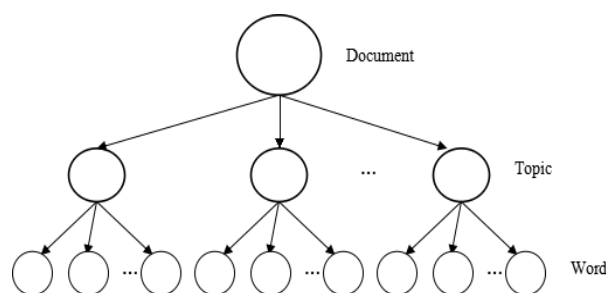


FIGURE 1 The structure of LDA

The main idea of LDA model is that document is represented as the probability distribution of topics, and each topic is represented as the probability distribution of words [3], which is shown in Figure 2.

*Corresponding author e-mail: 874036730@qq.com

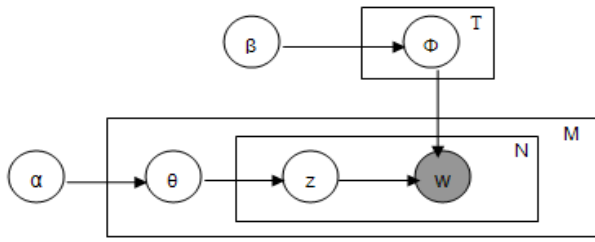


FIGURE 2 LDA model

The meaning of each symbol in LDA model diagram is shown in Table 1.

TABLE 1 The meaning of each symbol

Symbol	Meaning	Symbol	Meaning
α	Super parameter of θ	w	Word
β	Super parameter of Φ	N	The number of words
θ	Document-topic probability distribution	T	The number of topics
ϕ	Topic-word probability distribution	M	The number of documents
z	Word probability distribution		

Assuming that the document set $D = \{d_1, d_2, \dots, d_m\}$, one document of D is $d = \{w_1, w_2, \dots, w_n\}$, the number of topics of D is T , the probabilistic generative process of LDA model is as follows [4]:

- 1) For each topic t , a word multinomial distribution $\phi^{(t)}$ is obtained from the *Dirichlet*(β) distribution.
- 2) For each document d , a topic multinomial distribution θ_d is obtained from the *Dirichlet*(α) distribution.
- 3) For each word w_i in each document, extract a topic t from topic multinomial distribution θ_d , and extract a word as w_i from word multinomial distribution $\phi^{(t)}$ of the topic

From the LDA topic model, we can see that the conditional probability of the word w in a document can be calculated as Equation (1):

$$p(w | d, \hat{\theta}, \hat{\phi}) = \sum_{z=1}^T p(w | z, \phi) p(z | \hat{\theta}, d), \quad (1)$$

where z is the topic corresponding to w , T is the number of topics, $\hat{\theta}$ and $\hat{\phi}$ is the prior estimate of parameters θ and ϕ respectively.

3 Steps of SMS text similarity analysis based on LDA

SMS text is consisted of a series of independent short text essentially, which contains some special symbols to represent the interaction between user's behaviour and topics. In this paper, we take Chinese SMS text as an example to introduce the process of SMS text similarity calculation.

The analysis process of Chinese SMS text similarity is realized by text pretreatment, feature selection, text representation, LDA modelling and semantic similarity calculation, all together 5 steps, the specific

implementation steps and key problems will be described in followed each section. The analysis steps are as shown in Figure 3.

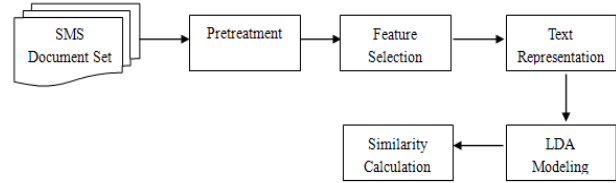


FIGURE 3 Steps of SMS text similarity analysis based on LDA

3.1 PRETREATMENT

Text pretreatment mainly refers to word segmentation, word tagging, and remove stop words, and so on. Chinese SMS text is a very short (less than 74 characters) and narrative essay information. We can remove the very small or irrelevant words with the topic, and use the named entity recognition technology to process personal name, place name, organization name, telephone number and other special information of SMS text, when performing feature selection, we can ignore these special information to avoid affecting the result of text clustering. In addition, the synonyms or near synonyms words are combined and expressed by the unified words.

3.2 FEATURE SELECTION

Feature selection can improve similarity precision for SMS text, and not all the words in SMS text play significant roles on similarity calculation. The words prevalent in all SMS text have no good distinguishing ability and have no great help on similarity calculation, even can affect the SMS text similarity precision, so these words should be removed.

In this paper, we select a number of words which is nouns or verbs in each SMS text as keywords to represent text, thus reduce the dimension of text feature vector as much as possible.

The ordinary feature selection methods include mutual information (MI), information gain (IG), the weight of evidence for text (WET), CHI statistic (CHI), and expected cross entropy (ECE), etc [5]. This paper uses MI method which is good at Chinese text to extract feature words. Assuming that there is a word w_i and category c , the mutual information between w_i and c can be defined as Equation (2):

$$MI(w_i, c) = \log \frac{p(w_i \cap c)}{p(w_i)p(c)}, \quad (2)$$

where $p(w_i \cap c)$ represents the probability that the word w_i and the category c appears simultaneously, $p(w_i)$ represents the probability that w_i appears and $p(c)$ represents the probability that c appears.

3.3 TEXT REPRESENTATION

This paper uses vector space model (VSM) to represent Chinese SMS text, which can be expressed as Equation (3):

$$D_{mi} = \{w_1, w_2, \dots, w_n\}. \quad (3)$$

The feature's weight is calculated by using the famous TF-IDF equation, which is expressed as Equation (4):

$$w_i(d) = \frac{tf_i(d) \log(N/n_i + 0.1)}{\sum_{i=1}^n t(f_i(d))^2 \times \log^2(N/n_i + 0.1)}, \quad (4)$$

where $w_i(d)$ represents the weight of word i in document d , and $tf_i(d)$ represents the word frequency of i in document d , N represents the total number of training texts, n_i represents the number of text which contains i and i appears in training texts. The denominator is the normalization factor.

3.3 LDA MODELLING

In the process of LDA modelling, the parameters of model are required to be estimated, the common estimation method are variational Bayesian inference, expectation propagation algorithm and Collapsed Gibbs sampling. Gibbs sampling algorithm is the most common method, it is easy to understand and implementation, and extract topics efficiently from large scale document set. In this paper, the method used for parameter estimation is Gibbs sampling.

Two matrixes can be obtained from Gibbs sampling process: topic-word matrix and document-topic matrix, namely θ and φ [6], the calculation method is as shown follows Equations (5) and (6):

$$\theta_d = \frac{n_j^m + \alpha}{n_d^m + T\alpha}, \quad (5)$$

$$\theta_d = \frac{n_j^w + \beta}{n_j^w + W\beta}, \quad (6)$$

where n_j^m represents the number of words that document d_m assign to topic j , n_d^m represents the number of words which have been assigned to topic in d_m , n_j^w represents the frequency that word w assign to topic j , n_j^w represents the number of words which have been assigned to topic j .

In topic-word matrix, each row corresponds to feature word of the feature list, each column corresponds to each topic, matrix elements value represent the number of times that the feature word is assigned to corresponding topic. In document-topic matrix, each row corresponds to a SMS text message of data set, each column corresponds to each

topic, matrix elements value represents the number of times that feature words of SMS text is assigned to a particular topic.

3.5 TEXT SIMILARITY CALCULATION

Different models usually correspond to different similarity calculation method. This paper uses JS (Jensen-Shannon) distance formula which can measure the distance of probability distribution to calculate the similarity between two documents [7]. We obtain document-topic probability distribution and topic-word probability distribution by constructing LDA model. Therefore, the similarity calculation of two documents can be realized by computing the corresponding topic probability distribution. The distance of vector $p = (p_1, p_2, \dots, p_k)$ to $q = (q_1, q_2, \dots, q_k)$ is computed as follows Equation (7):

$$D_{js}(p, q) = \frac{1}{2} \left[D_{KL} \left(p, \frac{p+q}{2} \right) + D_{KL} \left(q, \frac{p+q}{2} \right) \right], \quad (7)$$

where $D_{KL}(p, q) = \sum_{j=1}^T p_j \ln \frac{p_j}{q_j}$. p and q represents topic probability distribution respectively.

4 Experiments and analysis

To verify the proposed SMS text similarity calculation method in this paper, we use the thought of single-pass incremental clustering to perform experiment on text clustering. The basic idea of text clustering is [8]: preset a clustering threshold u , sequentially process the input SMS text, and calculate the similarity between the new text and identified topic clusters, if the similarity value is greater than the threshold u , the new text is added to its maximum similarity topic cluster, otherwise, the SMS text is created as a seed topic.

The experimental evaluation indexes adopt F-measure to measure the text similarity. F-measure is a balance index, which combine precision rate with recall rate in information retrieval, F-measure is greater, the clustering effect is better.

Precision rate and recall rate are calculated as follows Equations (8) and (9):

$$p(i, j) = \frac{n_{ij}}{n_j}, \quad (8)$$

$$r(i, j) = \frac{n_{ij}}{n_i}, \quad (9)$$

where n_i represents the number of documents contained in the category i , n_j represents the number of documents contained in the clustering j , n_{ij} represents the number of documents where the clustering j belongs to the category i .

F-measure is defined as follows Equation (10):

$$F(i, j) = \frac{2p(i, j)r(i, j)}{p(i, j) + r(i, j)} \tag{10}$$

The F-measure of global clustering is defined as follows Equation (11):

$$F = \sum_i \frac{n_i}{n} \max\{F(i, j)\} \tag{11}$$

4.1 EXPERIMENTAL DATA

This experiment selected five categories of Chinese junk messages as test data, they are lottery, advertisement, service charge, invoice and eroticism, a total of 1000 SMS text messages, we annotated these messages manually.

4.2 EXPERIMENTAL PROCEDURE

The specific experimental procedure is as follows:

1) SMS document set was pre-processed by using ITCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) to execute word segmentation, remove stop words, and so on. We obtained the text vector matrix.

2) LDA was introduced to model the text vector, and used Gibbs sampling method to estimate parameter. We obtained two matrixes: the document-topic matrix and topic- word matrix.

3) The similarity matrix was obtained by using JS distance formula to compute the similarity of SMS text messages.

4) Clustered the SMS text based on the similarity matrix by using single-pass incremental clustering algorithm, and analyzed the clustering results.

4.3 EXPERIMENTAL ANALYSIS

According to the experience value, we set $\alpha = 50/T$ and $\beta = 0.01$, the number of topics T directly affects the precision of LDA model, which will affect the accuracy of clustering results, so we determined the value of T by experiment. The F -measure is higher, the clustering effect is better. Different number of topics generates different F -measure of the clustering, which is as shown in Figure 4.

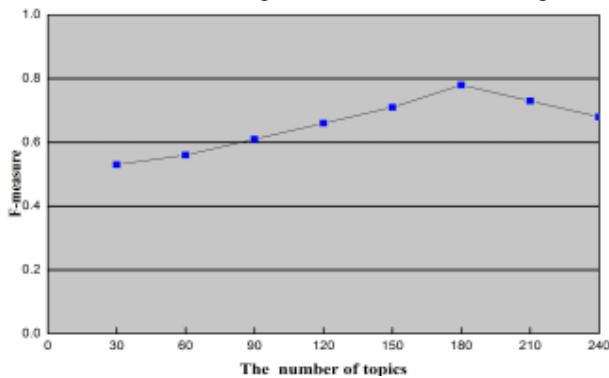


FIGURE 4 Relationship between the number of topics and F-measure

Here the abscissa represents the number of topics, the ordinate represents the size of the F-measure. The Figure 4 reflects the change of F-measure with the different number of topics. As can be seen from Figure 4, the F-measure is highest when $T = 180$, so the number of topics is determined as 180 in this paper.

The experimental statistic found that the similarity of SMS text under the same category of experimental data is commonly above 0.49, and the similarity of SMS under the different categories is generally below 0.35. Therefore, the experiment set the range of clustering threshold u , where $u \in [0.35, 0.49]$. Experiments were carried out 8 times, and used the TDT standard as evaluation index [9], system performance was evaluated by using the macro average of detection cost in 5 categories SMS, The normalized detection overhead changed along with the change of cluster threshold, which is as shown in Figure 5.

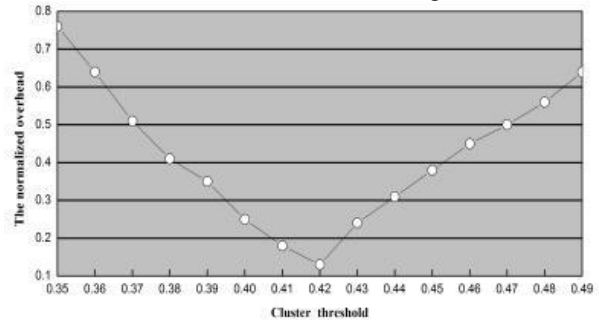


FIGURE 5 Determine the optimal cluster threshold

Here the abscissa represents the preset clustering threshold, the ordinate represents the system normalized overhead. As can be seen from Figure 5, when the threshold is taken 0.42, the system normalized overhead is the minimum, and its value is 0.125, so determine the cluster threshold u is 0.42 in this paper.

To verify the effectiveness of our proposed similarity calculation method based on LDA, we adopt the statistical method based on TF-IDF (referred to as statistics-based) and the semantic method based on Chinese HowNet dictionary (referred to as dictionary-based) as comparative experiments, and compare their clustering effects by single-pass incremental clustering algorithm, Experimental comparative results of three different methods are as shown in Table 2.

TABLE 2 Experimental results of three different methods

Experimental method	Precision (%)	Recall (%)	F-measure (%)
Statistics-based	81.35	81.68	81.51
Dictionary-based	84.24	85.19	84.71
LDA-based	87.87	88.37	88.12

From Table 2, we can see that this proposed method achieves better effects compare with other two methods.

5 Conclusions

In view of the defect of existed statistics-based method and dictionary-based method, this paper takes full advantage

of topic model to calculate the text similarity of Chinese SMS, uses LDA to construct the SMS text topic space, comprehensively considers the similarity from semantic similarity of words and word probability distribution in the text, enhances the vector representation of documents, at the same time, greatly reduces the dimension of the document, speeds the computing speed, thus improves the clustering effect. Experimental results show the effectiveness of this method.




In the next step work, we still need to improve the accuracy of document-topic probabilistic distribution and topic-word probabilistic distribution of LDA model, which make an important impact on text clustering.

References

- [1] Liu J L, Song L Y, Fan Y H 2012 Study of Chinese SMS Text Similarity Based on Semantic Information *Computer Engineering* **38**(13) 58-62 (in Chinese)
- [2] Blei D M, Jordan M 2003 Modelling annotated data *Proceedings of 26th Annual International SIGIR Conference on Research and Development in Information Retrieval SIGIR'03 ACM* 127-34
- [3] Blei D M, Ng A Y, Jordan M I 2003 Latent Dirichlet Allocation *The Journal of Machine Learning Research* **3**(1) 993-1022
- [4] Tan C F 2013 Short Text Classification Based on LDA and SVM *International Journal of Applied Mathematics and Statistics (IJAMS)* **51**(22) 205-14
- [5] Yahya W B, Ulm K, Fahrmeier L, Hapfelmeier A 2011 A sequential feature selection and prediction method in microarray studies *International Journal of Artificial Intelligence* **6**(11) 19-47
- [6] Quan X, Liu G, Lu Z, Ni X, Liu W 2010 Short Text Similarity Based on Probabilistic Topics *Knowledge and Information Systems* **25**(3) 473-91
- [7] Griffiths T L, Steyvers M 2004 Finding scientific topics. *Proceedings of 3th National Academy of Sciences* **101** 5228-5235
- [8] Zhao A H, Liu P Y, Zheng Y 2013 Subtopic Division in News Topic Based on Latent Dirichlet Allocation *Journal of Chinese Computer Systems* **34**(4) 732-7 (in Chinese)
- [9] The 2004 topic detection and tracking (TDT2004) task definition and evaluation plan 2004 <http://www.nist.gov>

Acknowledgements

This work was supported by Key University Science Research Project of Anhui Province (No.KJ2014A250) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2014YKF41) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2013YKF19) and Software engineering projects (No.2013zytz074) and The training system and training platform construction based on engineering training and innovation ability of computer professional (No.2013cgtg032).

Authors	
	<p>Chengfang Tan, born in February, 1981, China</p> <p>Current position grades: Researcher at Intelligent Information Processing Lab, Suzhou university, China. University studies: Master degree in education technology from Nanjing Normal University, China in 2007. Scientific interests: information retrieval, sentiment analysis and text mining.</p>
	<p>Caiyin Wang, born in September, 1978, China</p> <p>Current position grades: Researcher at Intelligent Information Processing Lab, Suzhou University, China. University studies: Master degree in computer science and technology at Hefei University of Technology, China in 2009. Scientific interests: P2P and information retrieval.</p>
	<p>Lin Cui, born in August, 1979, China</p> <p>Current position grades: Researcher at Intelligent Information Processing Lab, Suzhou university, China. University studies: Master degree in computer science and technology from Hefei University of Technology, China in 2008. Scientific interests: information retrieval and Semantic Web.</p>